# Better Decision Making for New York City Taxi Drivers

Kefu Zhu, Chunlei Zhou

## 1. Introduction

As widely recognized symbols of the New York City, yellow and green taxis have a profound history since 1890s. Taxicabs provide both tons of employment positions and convenient commute options for the people in New York City under the supervision of Taxi and Limousine Commission (TLC). In this project, we will explore the the data source from TLC[1] and study how can a taxi driver make better decisions, as well as the customer profile of taxi passengers within the New York City. In this project, we are interested to answer three questions using the TLC taxi data. Where should the drivers go for pickups if they want higher tip earnings? Will the effect of weather and time influence the income of taxi drivers. And lastly, what are the intrinsic properties of different customer groups for both yellow and green taxi?

## 2. Related Works

As TLC made the taxi trip records available to the public, a wide range of analysis has been done based on the data. Previous studies using the New York City Taxi data including how the trend changes among different types of taxis across time and area [1], what are the factors that influences the ride duration [2], what is the average speed of a trip at a particular hour of the day [3], where are the hot pickup/dropoff areas in the New York City [4] and many other interesting topics [5][6].

## 3. Methodology

[1] TLC Trip Record Data: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

The data is programmatically downloaded and loaded into python from the TLC website, which provides different datafile in CSV format every month from January 2009 to June 2019. In this project, due to the limited time and computing resources, we only used the 2018 data for yellow taxi and green taxi in both March (winter season) and August (summer season). The detailed dictionary of attributes in the dataset is provided on the TLC website[2].

To explore the customer profile of different types of taxi, we experimented with K-means, K-modes and DBSCAN. In addition, we also used hypothesis testing when investigating the impact of weather and time on driver's income.

## 4. Experiment

### 4.1 Tip in Different Pick Up Locations

To answer the question, where should the drivers go for pickups if they want higher tip earnings, we examined two metrics: the tip amount and tip as a percentage of total fare. The only location information in the taxi dataset is LocationID, which is the ID of different areas within the New York City.

[2] Data Dictionary: https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

```
Green Taxi: Tip_Percentage Median - Pickup_Location Relationship
Best Location: 2.0
```

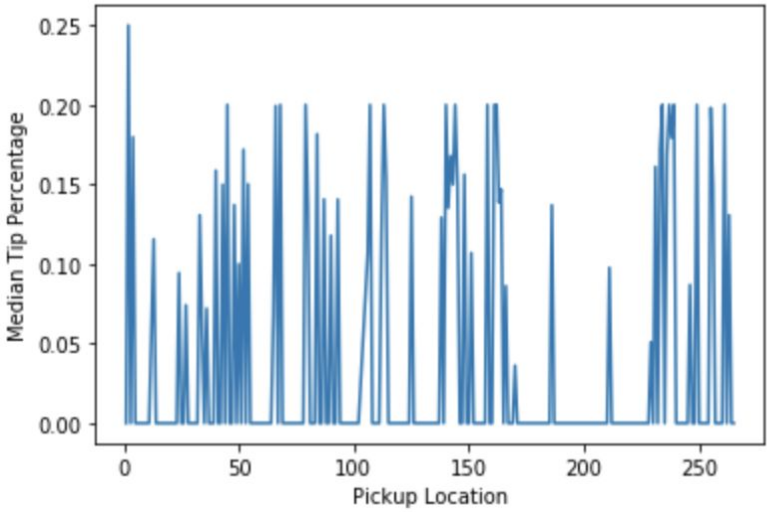


Figure 1. Median Tip as a Percentage of Total Fare in Different Pickup Location for Green Taxi

```
Yellow Taxi: Tip_Percentage Median - Pickup_Location Relationship
Best Location: 2.0
```

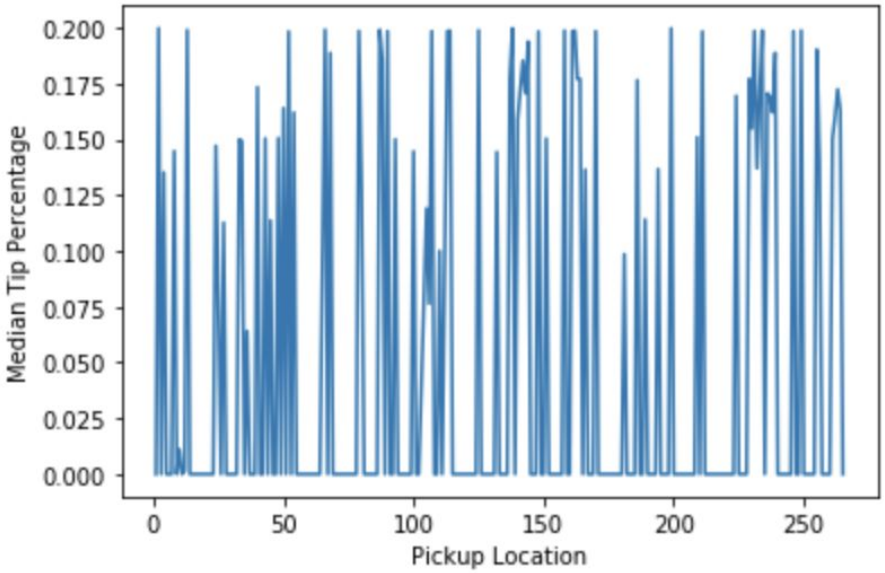


Figure 2. Median Tip as a Percentage of Total Fare in Different Pickup Location for Yellow Taxi

```
Green Taxi: Tip_Amount Median - Pickup_Location Relationship
Best Location: 84.0
```

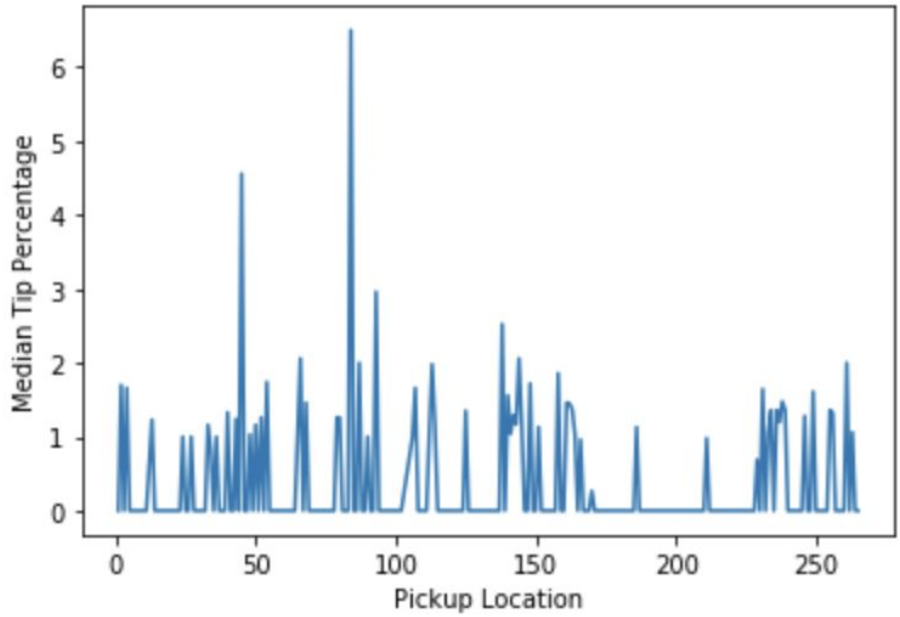


Figure 3. Median Tip Amount in Different Pickup Location for Yellow Taxi

```
Yellow Taxi: Tip_Amount Median - Pickup_Location Relationship
Best Location: 2.0
```

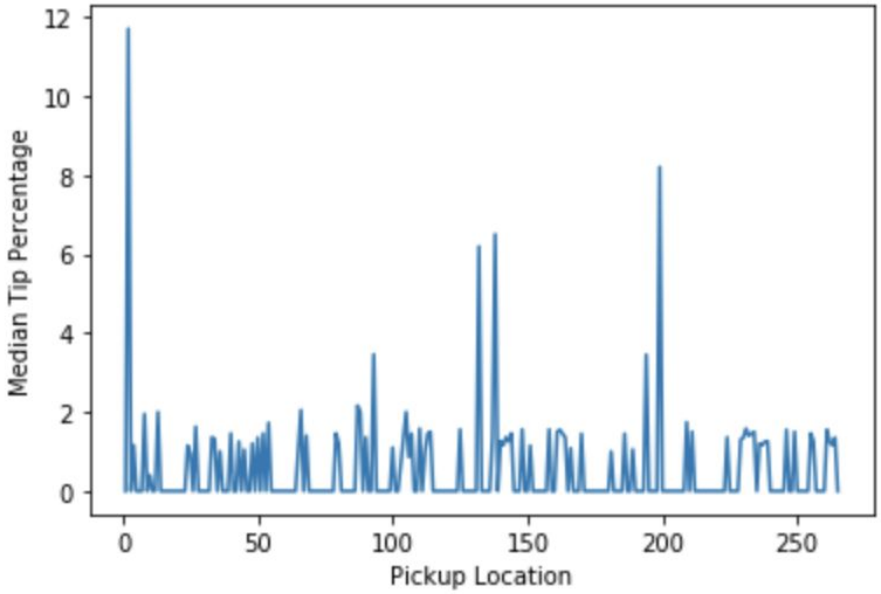


Figure 4. Median Tip Amount in Different Pickup Location for Yellow Taxi

Based on the analysis result shown above, we notice that among total of 257 regions, trips that pick up passengers in some regions tend to have higher tip amount as well as higher tip as a percentage of total fare.

| Top 10 Location for Yellow Taxi | | |
|---|---|---|
| Location | Median Tip Amount ($) | Median Tip Percentage |
| 199 | 8.2 | 0.199708 |
| 138 | 6.21 | 0.199693 |
| 10 | 6 | 0.126863 |
| 2 | 5.95 | 0.13544 |
| 93 | 5 | 0.151515 |
| 132 | 5 | 0.119536 |
| 215 | 3 | 0.085383 |
| 194 | 2.4 | 0.112971 |
| 87 | 2.06 | 0.198795 |
| 13 | 2 | 0.198864 |

Figure 5

| Top 10 Location for Green Taxi | | |
|---|---|---|
| Location | Median Tip Amount ($) | Median Tip Percentage |
| 84 | 6.5 | 0.181633 |
| 45 | 4.56 | 0.2 |
| 93 | 2.965 | 0.141094 |
| 138 | 2.56 | 0.131579 |
| 66 | 2.06 | 0.199275 |
| 144 | 2.06 | 0.2 |
| 87 | 2 | 0.140684 |
| 261 | 2 | 0.2 |
| 133 | 1.98 | 0.2 |
| 158 | 1.86 | 0.2 |

Figure 6

| Top 10 Location for Yellow Taxi | | |
|---|---|---|
| Location | Median Tip Amount ($) | Median Tip Percentage |
| 199 | 8.2 | 0.199708 |
| 138 | 6.21 | 0.199693 |
| 52 | 1.76 | 0.199074 |
| 13 | 2 | 0.198864 |
| 87 | 2.06 | 0.198795 |
| 125 | 1.56 | 0.198795 |
| 158 | 1.56 | 0.19863 |
| 162 | 1.54 | 0.19863 |
| 234 | 1.5 | 0.19863 |
| 249 | 1.46 | 0.19863 |

Figure 7

| Top 10 Location for Green Taxi | | |
|---|---|---|
| Location | Median Tip Amount ($) | Median Tip Percentage |
| 2 | 1.7 | 0.25 |
| 45 | 4.56 | 0.2 |
| 68 | 1.46 | 0.2 |
| 79 | 1.26 | 0.2 |
| 113 | 1.98 | 0.2 |
| 140 | 1.56 | 0.2 |
| 144 | 2.06 | 0.2 |
| 158 | 1.86 | 0.2 |
| 161 | 1.46 | 0.2 |
| 162 | 1.45 | 0.2 |

Figure 8

From the summary above, we also find by looking at different metrics, the top 10 locations for high tip can vary. However, some regions remain in the list even when switching the evaluation metric. Location 13, 87, 138 and 199 (highlighted in yellow) are in the top 10 list for yellow taxi drivers, no matter which metric is used. Similarly, location 45, 144 and 158 for green taxi.

## 4.2 Drivers Income in Extreme Weather

We gathered weather warning information from the archived database maintained by Iowa State University. The data is originated from the National Weather Service (NWS) and it contains different levels of warnings issued by NWS for extreme weather events. Since we are interested in the relationship between extreme weather and income of taxi drivers in NYC, we only focussed on the events that have the most severe level of warning.

In 2018, NYC experienced 2 winter storms in March (03/06 - 03/08 and 03/20 - 03/22) and 2 severe thunderstorms in August (08/07 and 08/11). We proposed two metrics to measure the income of taxi driver, income per second ($/s) and total fare ($).

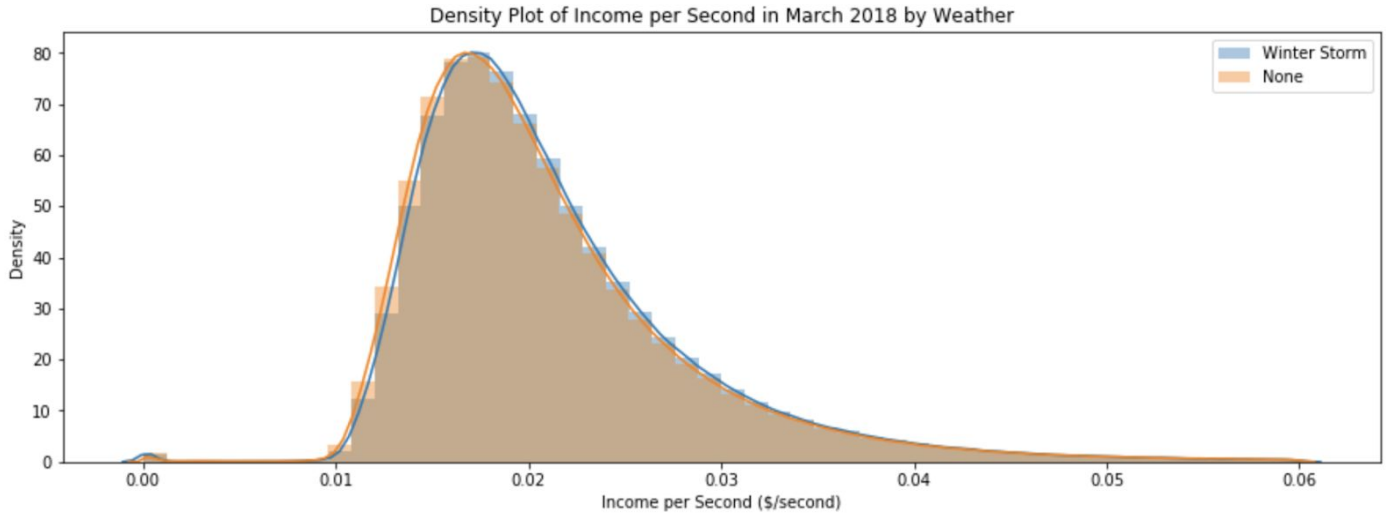


Figure 9. Density Plot of Income per Second in March 2018 for Yellow Taxi Drivers

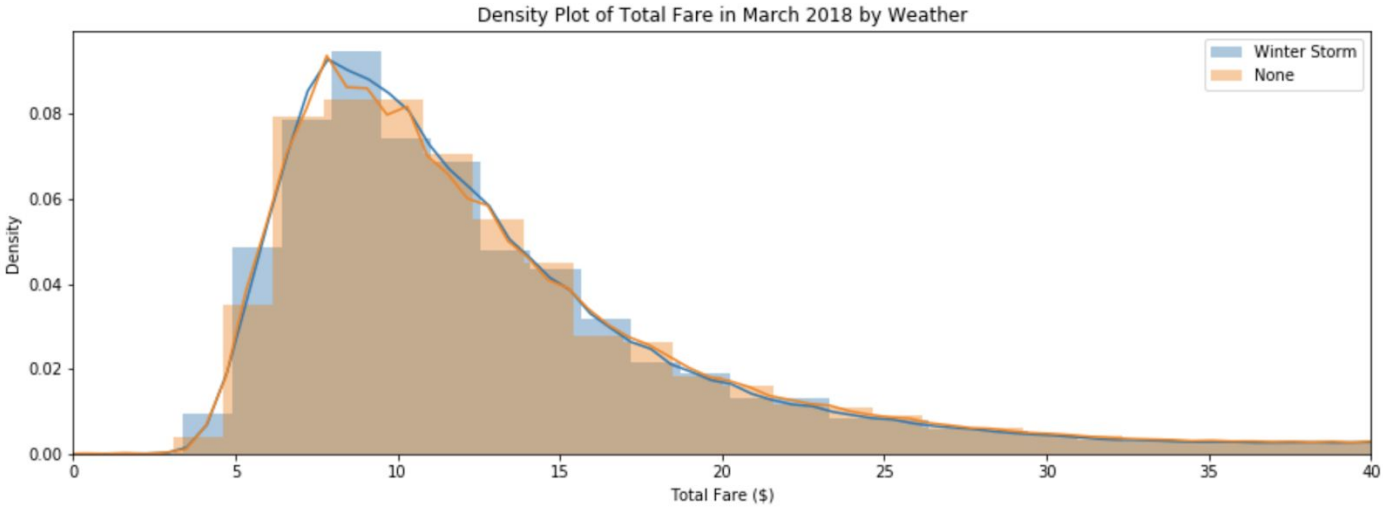


Figure 10. Density Plot of Total Fare in March 2018 for Yellow Taxi Drivers

After removing extreme values, as shown on the graphs (Figure 9, 10, 11 and 12), surprisingly, the income of yellow taxi drivers in NYC are not affected by either winter storm or severe thunderstorm no matter which metric is used to evaluate. The result for green taxi is also the same. Because we can barely see any difference from the density plot, we felt there is no point to proceed on doing the hypothesis testing.

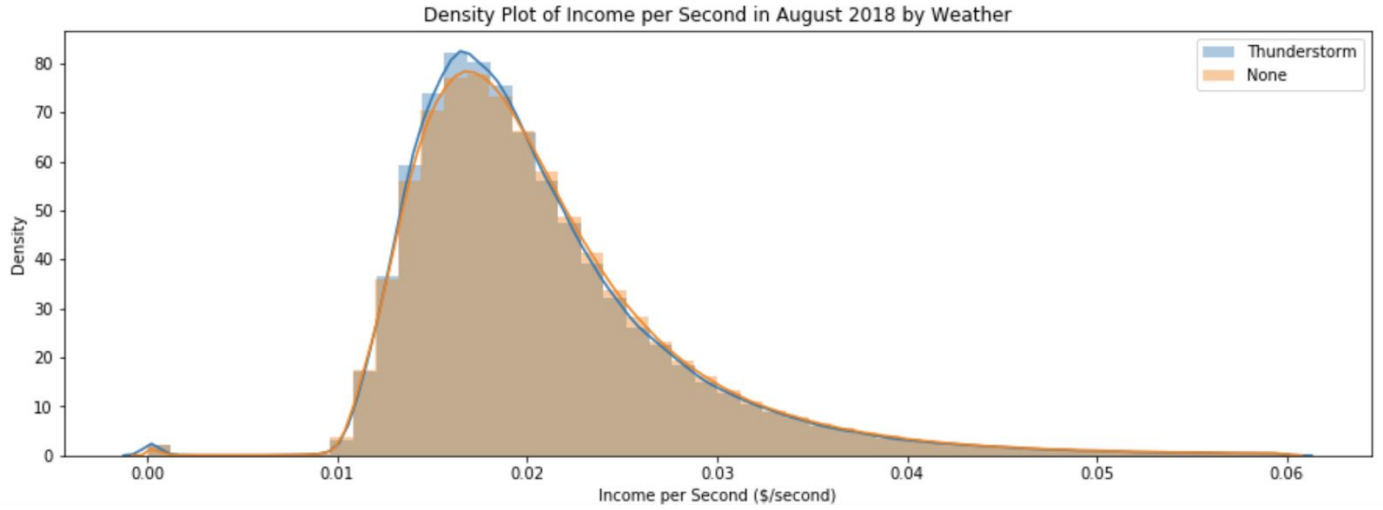


Figure 11. Density Plot of Income per Second in August 2018 for Yellow Taxi Drivers

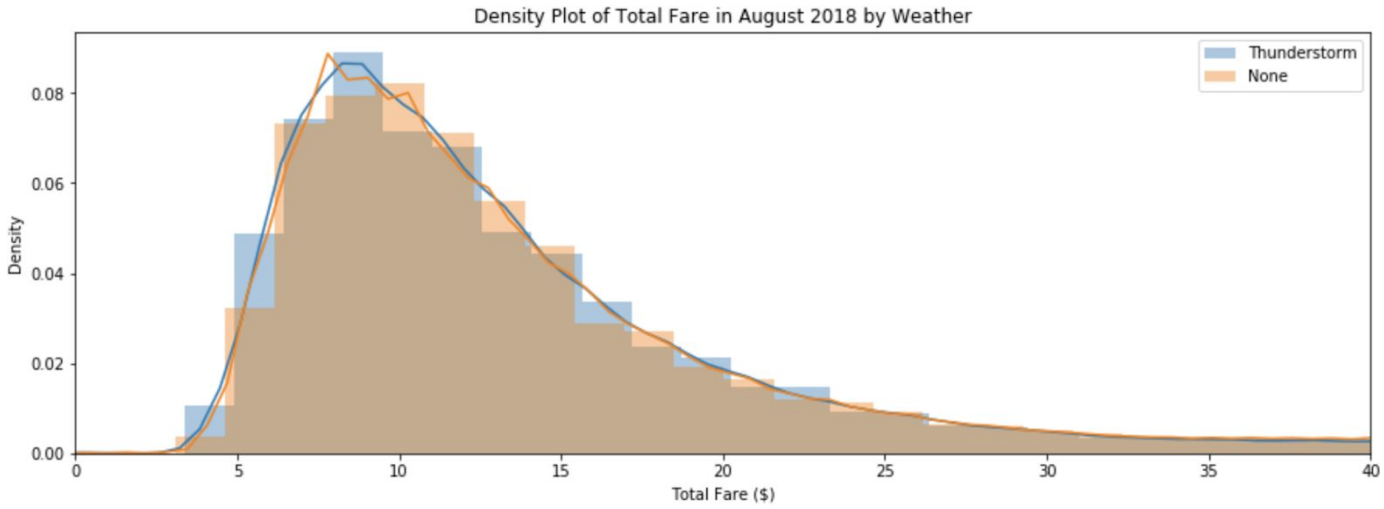


Figure 12. Density Plot of Total Fare in August 2018 for Yellow Taxi Drivers

## 4.3 Drivers Income in Different Times of a Day

The original pickup and dropoff time of each trip is recorded as timestamp. We aggregated the trips into five different time window for better analysis: morning (6am - 11am), noon (11am - 2pm), afternoon (2pm - 6pm), night (6pm - 12am) and late-night (12am - 6am). Similar to the analysis in Section 4.2, we used the same two metrics here.

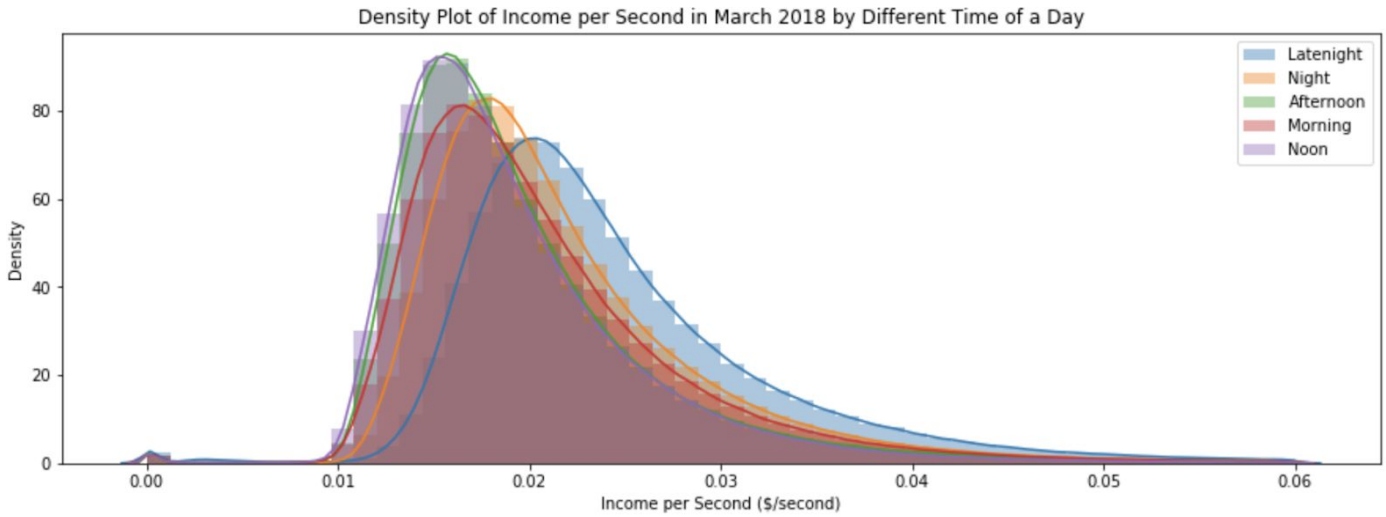


Figure 13. Density Plot of Income per Second in March 2018 for Yellow Taxi Drivers

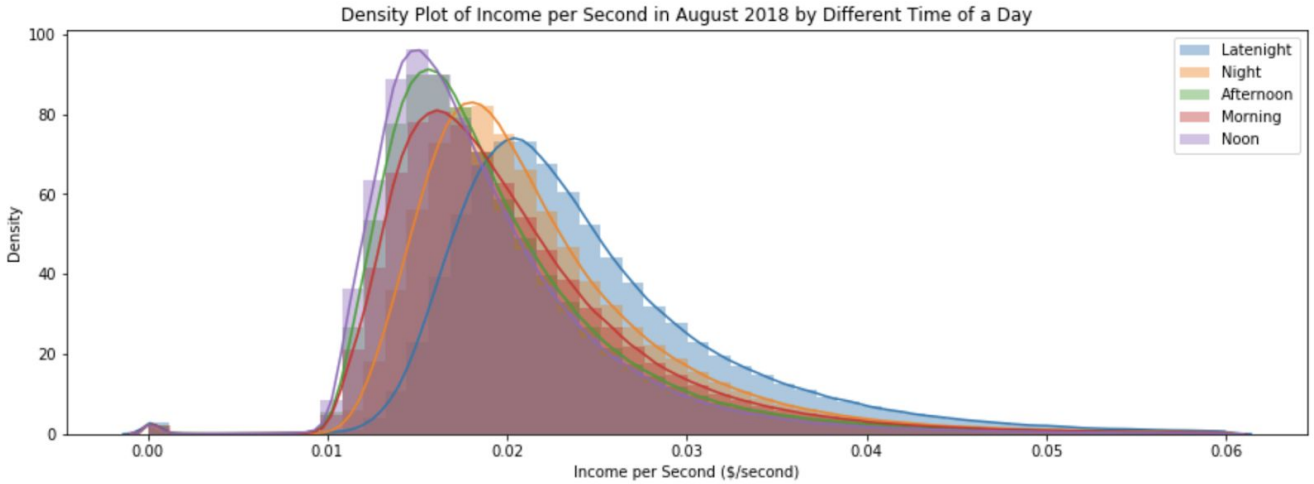


Figure 14. Density Plot of Income per Second in August 2018 for Yellow Taxi Drivers

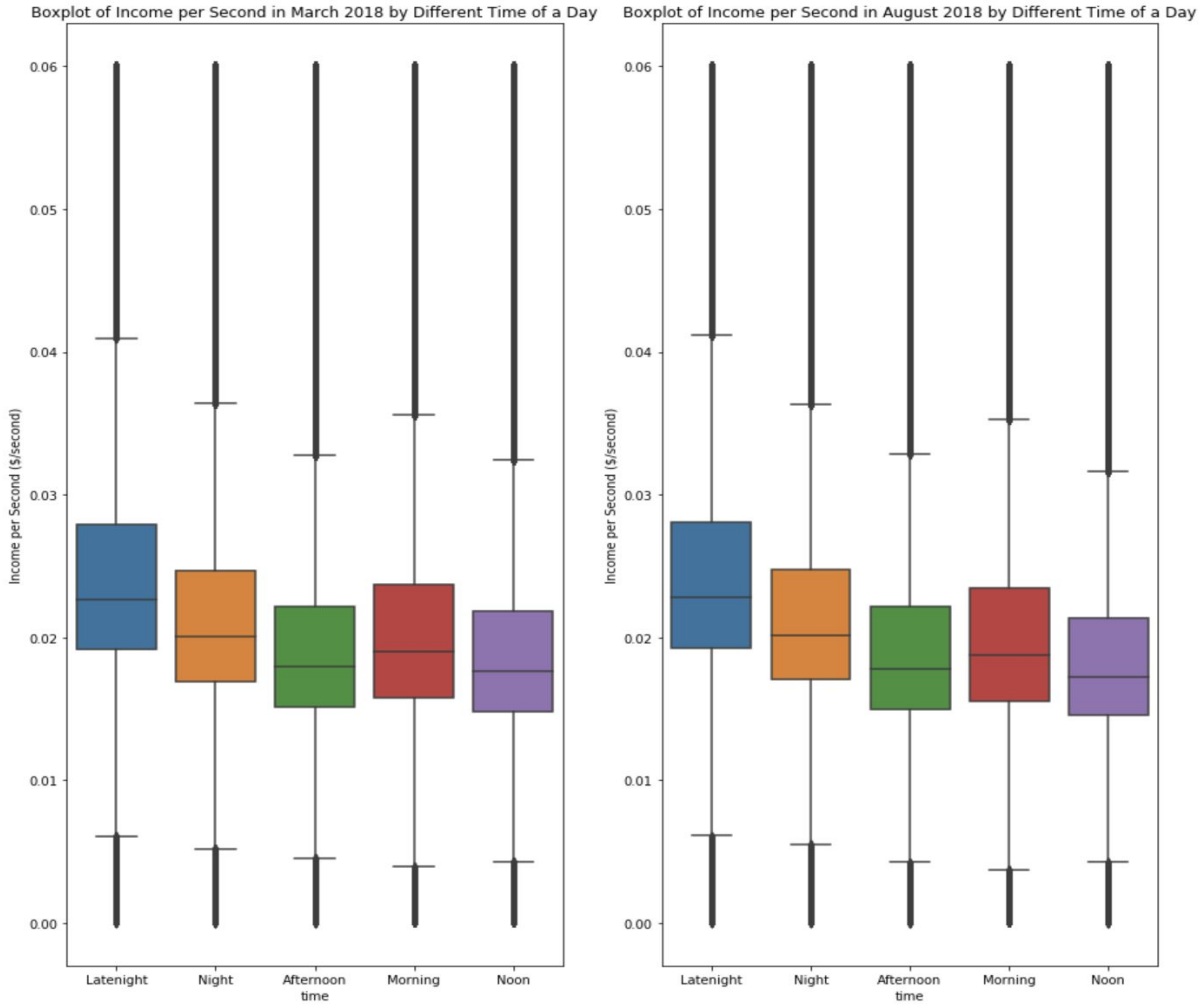


Figure 15. Box Plot of Income per Second in March and August 2018 for Yellow Taxi Drivers

Based on both the density plot and the box plot above, we do observe some differences in the income per second of yellow taxi driver at different time of a day. Generally speaking, during the night (6pm - 12am) and late-night (12am - 6am), taxi drivers earn higher income per second. Trips in the morning (6am - 11am) also tend to have higher return compared to trips in the noon (11am - 2pm) and afternoon (2pm - 6pm). However, if we evaluate the trips using the other metric, total fare amount, we end up with a different story.

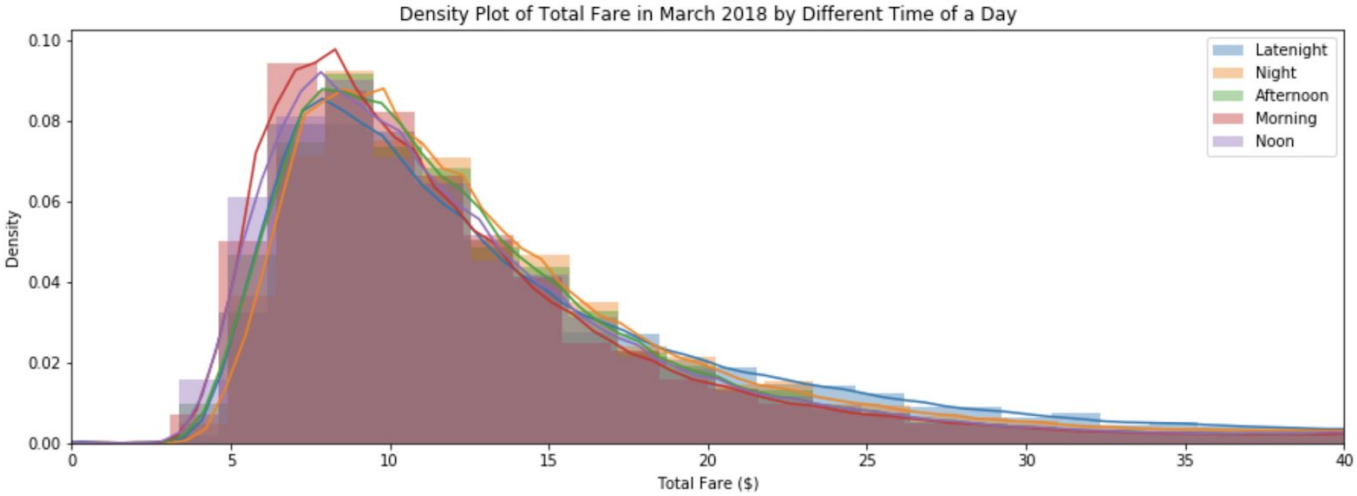


Figure 16. Density Plot of Total Fare in March 2018 for Yellow Taxi Drivers

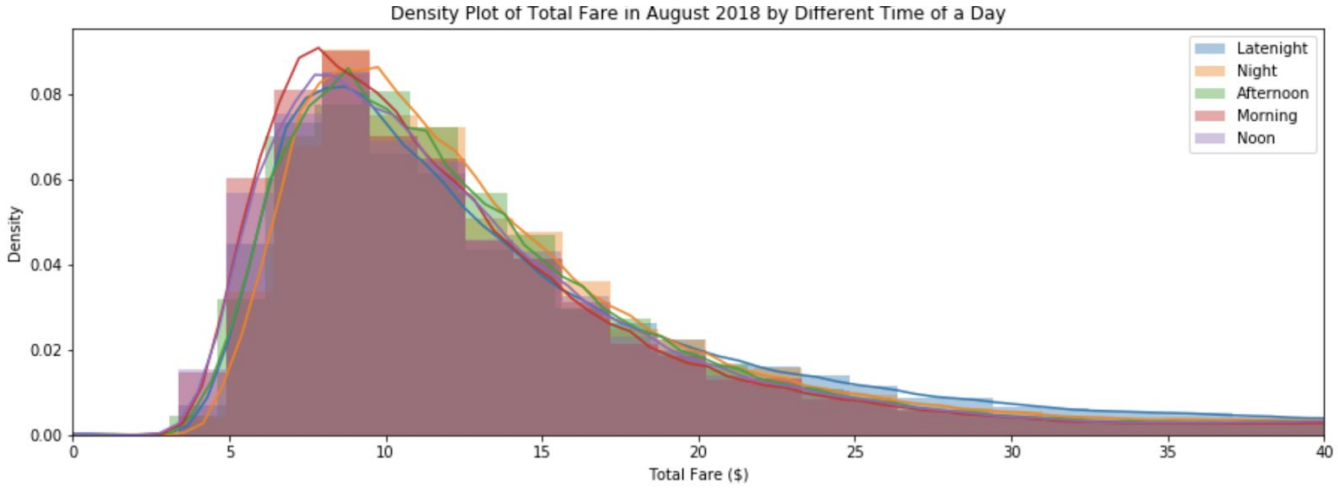


Figure 17. Density Plot of Total Fare in March 2018 for Yellow Taxi Drivers

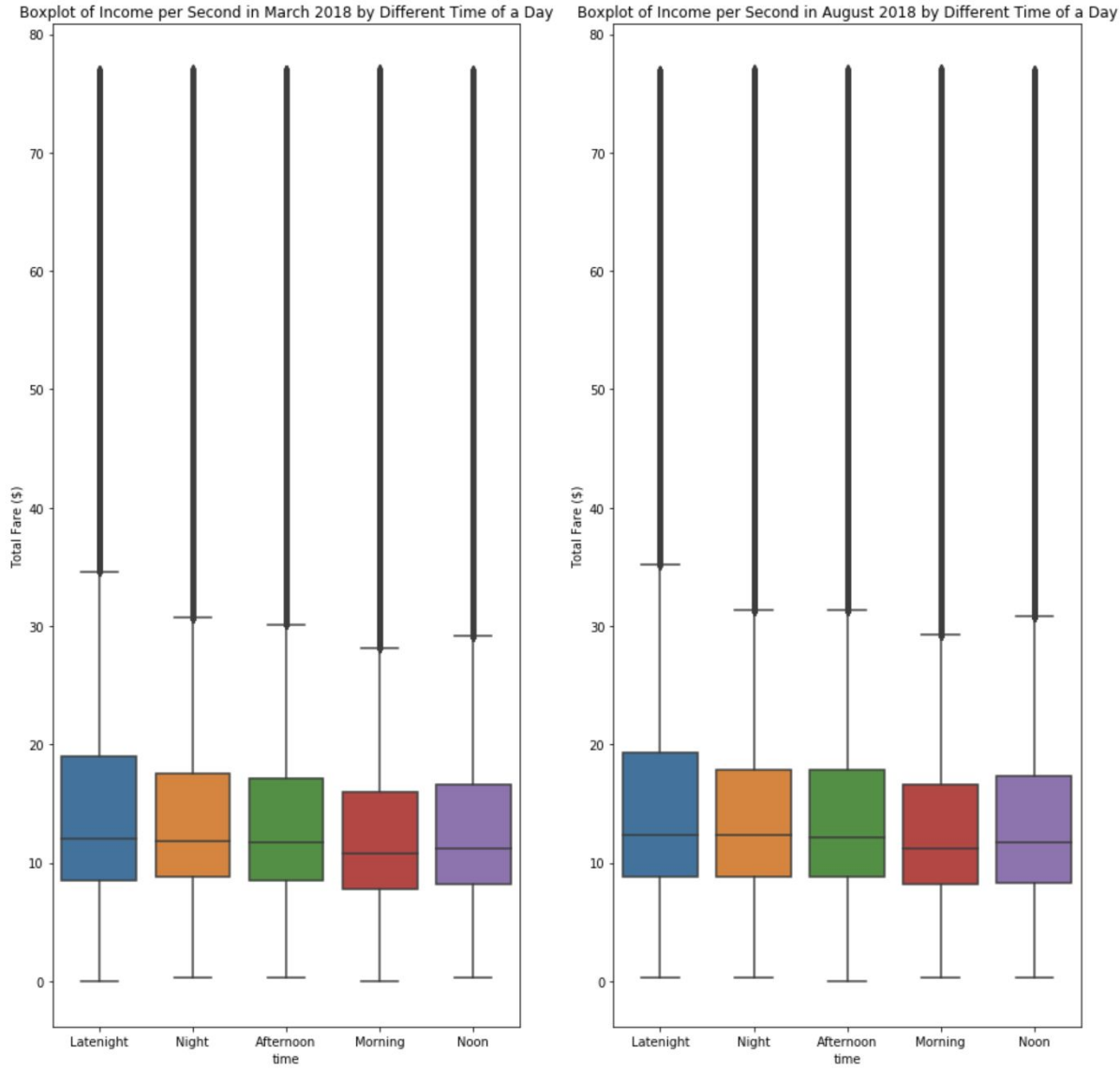


Figure 18. Box Plot of Total Fare in March and August 2018 for Yellow Taxi Drivers

In either the density plot or the box plot, it is hard to conclude that there is any significant difference among the total fare of trips in different time of a day for yellow taxi drivers. Furthermore, since the trips during night and late-night seem to yield higher return for yellow taxi drivers. We conducted hypothesis testing using pairwise t-test to compare the return from trips during the day (6am-8pm) and night (8pm-6am). In order to have balanced sample size, we

downsampled the trips originated during the day to the same number of trips during the night and performed the pairwise t-test. To make our analysis result more robust, we repeated the pairwise t-test five times with different downsampled results. All tests yield p-value that is really close to zero, which means the return from trips during the night is significantly higher than the return from trips during the day.

## 4.4 Customer Profile for Yellow Taxi and Green Taxi

Among all the works on the TLC taxi data, we have not seen any that study the underlying passenger profile. We experimented with two different clustering methods, DBSCAN and K-means/K-modes, aiming at capture any potential structures of NYC taxi passengers. Since the dataset contains both interesting numeric attributes and categorical attributes, we performed clustering twice for each method. For numeric attributes, we picked number of passengers, trip distance (in miles), total fare amount, tip amount and duration of the trip (in seconds). For categorical attributes, we only picked three: RatecodeID (Indicating which route the trip traveled), payment type, and time of a day.

### 4.4.1 K-means/K-modes

Because the original data is too large even for one month, we first sampled about 2 million trip records from it, then performed K-means clustering on numeric attributes and K-modes clustering on categorical attributes. We experimented with different values of K ($k = 3$, $k = 5$, $k = 10$). In order to compare with the DBSCAN results in the next section, we only show the clustering results of $k = 10$ here.

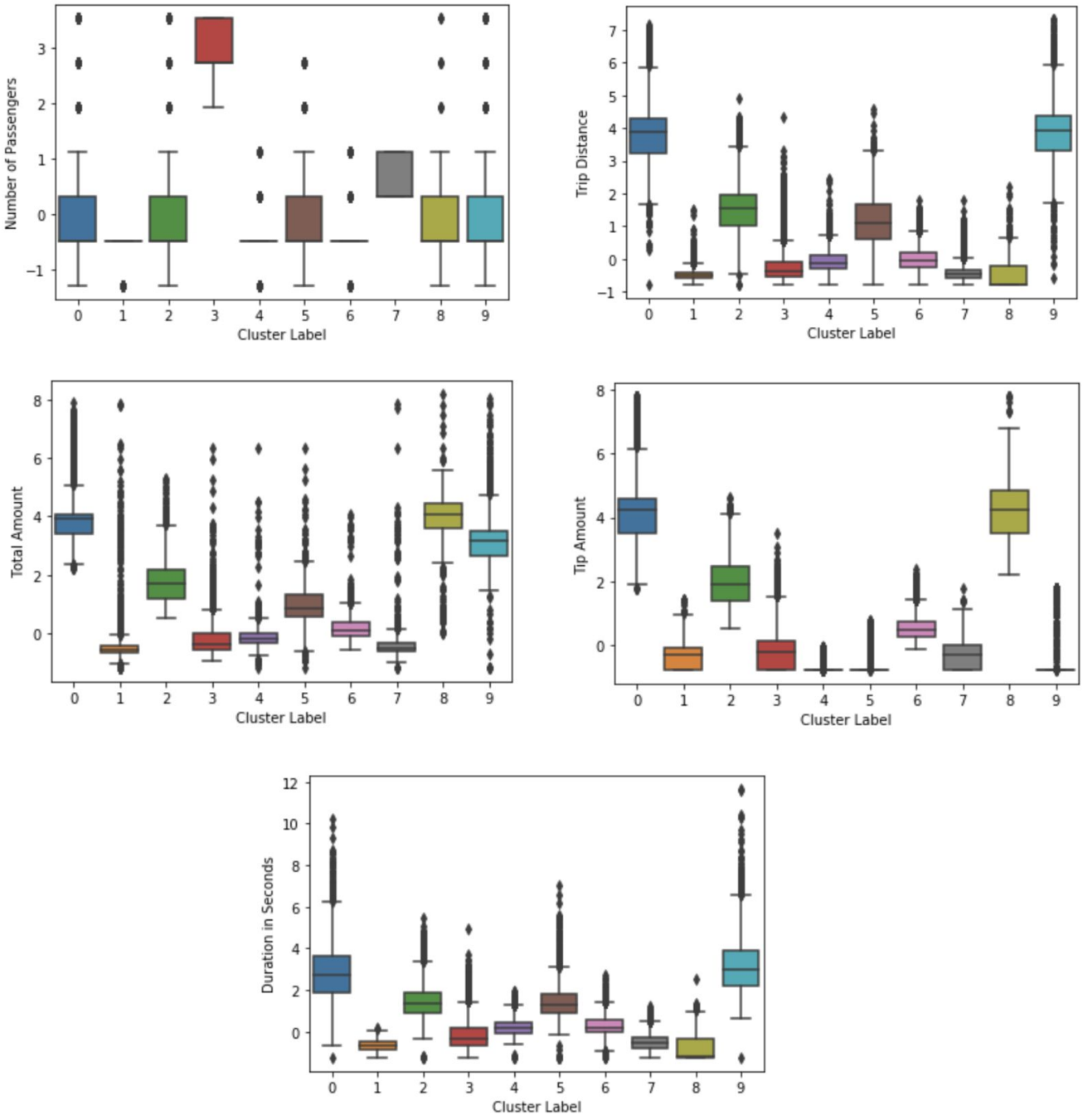


Figure 19. Numeric Attribute based K-means Results for Yellow Taxi

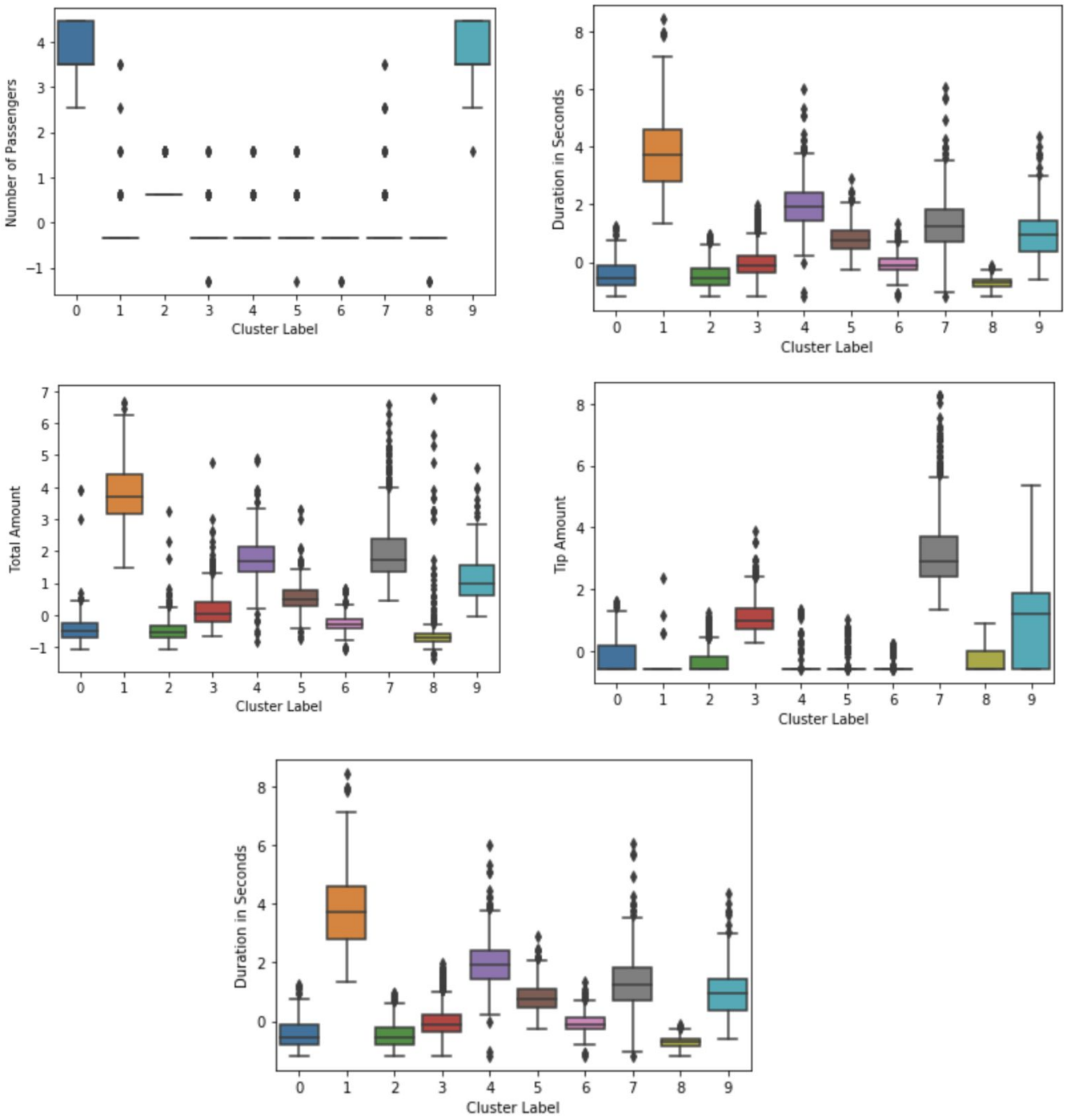


Figure 20. Numeric Attribute based K-means Results for Green Taxi

From the results above, we can see some interesting findings: trips that have more passengers give lower tip, and tip amount is not necessarily positively correlated with total fare amount.

To analyze the categorical attributes, instead of using K-means, we adapt the K-modes clustering method. In the graph below, time period is encoded as: 1 - morning, 2 - noon, 3 - afternoon, 4 night, 5 - latenight. Similarly, payment type is encoded as: 1 - Credit Card, 2 - No Charge, 3- Dispute, 4 - Unknown, 5 - Voided trip.

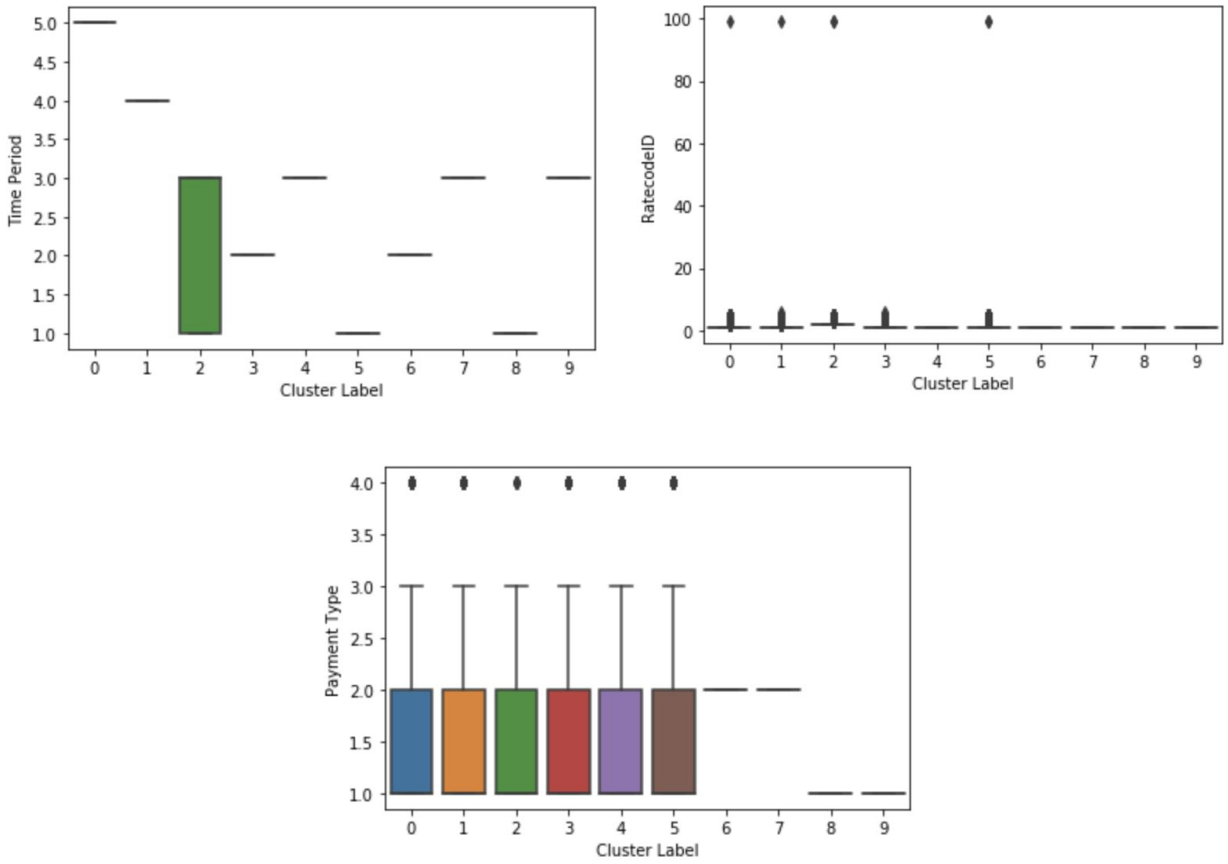


Figure 21. Categorical Attribute based K-modes Results for Yellow Taxi

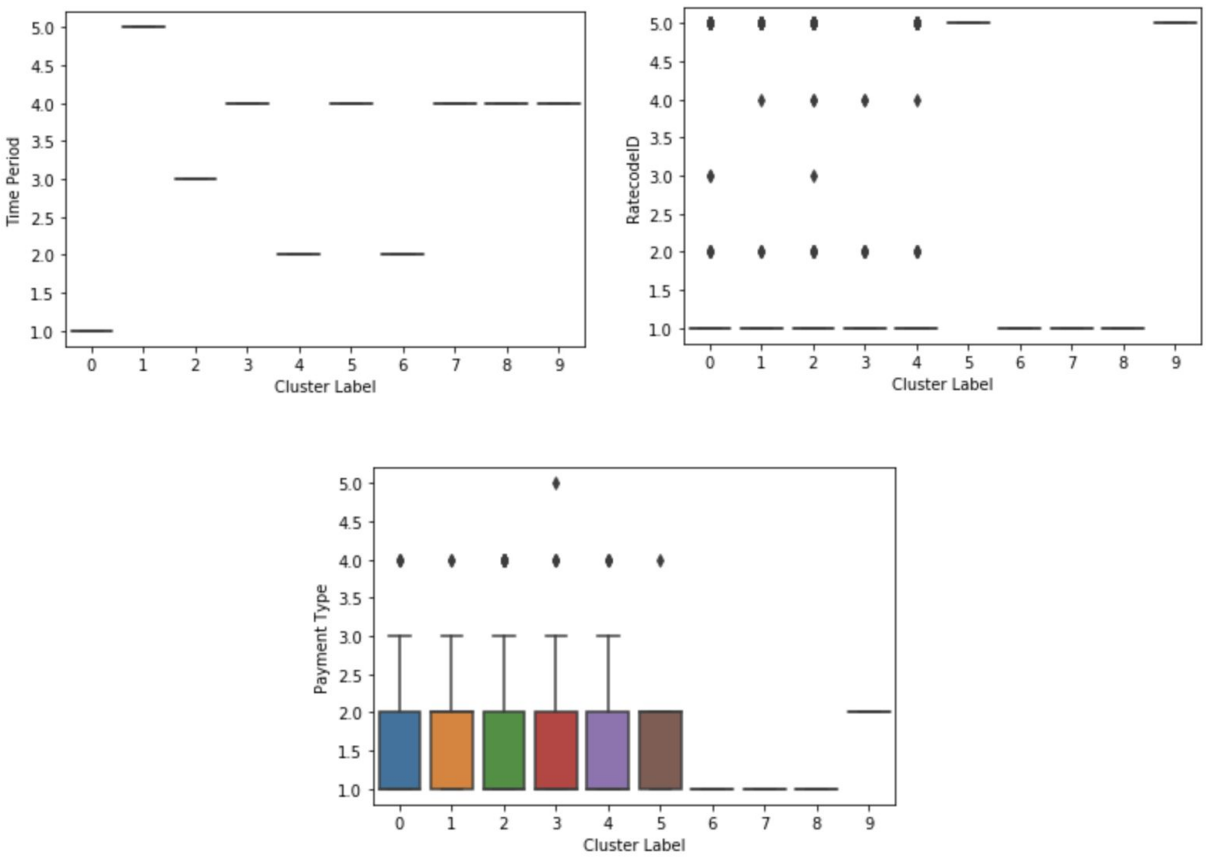


Figure 22. Categorical Attribute based K-modes Results for Green Taxi

Based on the clustering result, we can see K-modes mainly uses time period and Ratecode ID to separate the clusters. But the result in general does not reveal any interesting structure that can be rationally interpreted.

### 4.4.2 DBSCAN

The same data size is used when performing the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), with one given restriction: each cluster must contain at least 100 data points. In order to compute the distance properly when using categorical attributes, we one hot encoded the attributes and used the hamming distance as our distance metric.

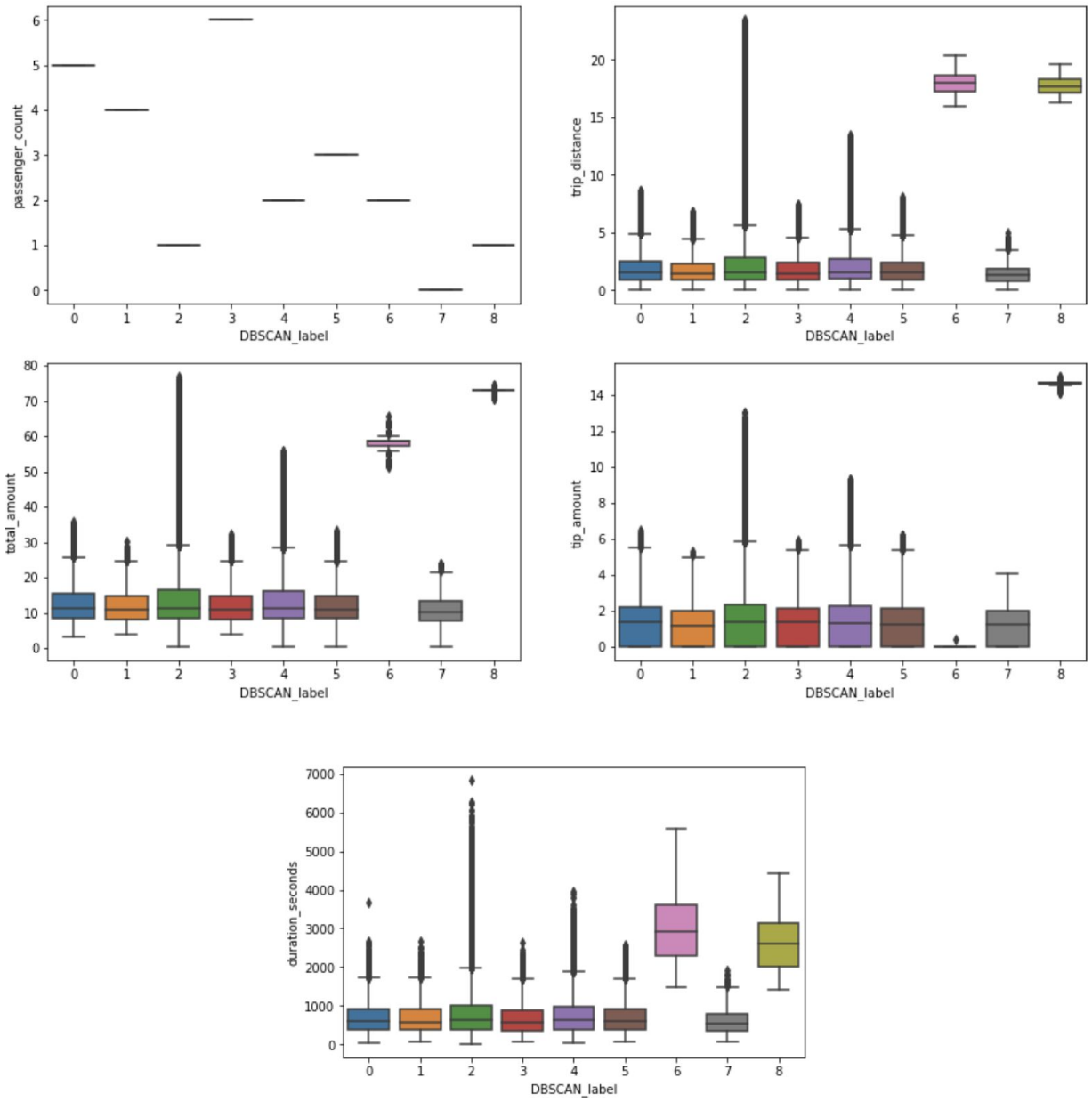


Figure 23. Numeric Attribute based DBSCAN Results for Yellow Taxi

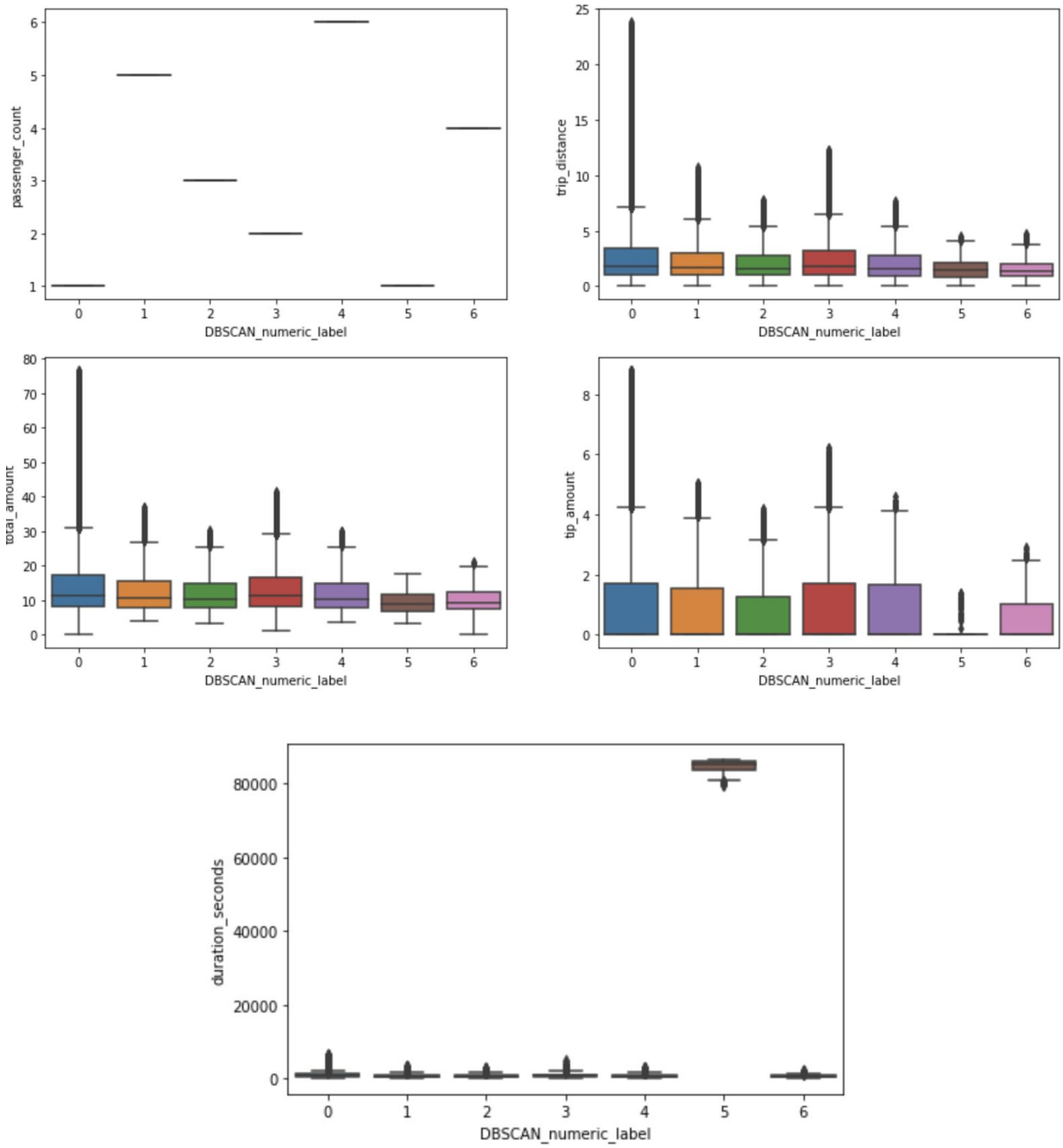


Figure 24. Numeric Attribute based DBSCAN Results for Green Taxi

Based on the clustering results using numeric attributes, we notice that DBSCAN uses number of passengers as main indicator for different clusters, for the rest of numeric attributes, only cluster

6 and cluster 8 behave differently for yellow taxi, which tend to have higher values in the selected numeric attributes compared to other clusters. For green taxi, the results basically provide no valuable information. Furthermore, when using categorical attributes as input, DBSCAN is not able to separate the data and outputs only one giant cluster.

## 5. Conclusion

Some regions in the New York city provide trips for the taxi driver that potentially have higher tips. As a NYC taxi driver, working on shift during the night (6pm - 12am) and late-night (12am - 6am) is likely to yield more income. Even though we may naturally expect trips during extreme weather (winter storm and severe thunderstorm) are different from normal trips. The analysis results indicate that there is no apparent difference between them from a perspective of driver's earnings. Finally, the clustering results from both K-means and DBSCAN reveal some underlying structures and interesting insights for the passengers of NYC taxi. Passenger groups with different size have different behaviors in trip distance (in miles), total fare amount, tip amount and duration of the trip (in seconds), but they do not necessarily differ in the going on a specific route or using distinctive payment types..

## 6. Bibliography

1. Todd W. Schneider. November 2015. Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. Retrieved from: https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

2. Kaggle. 2017. New York City Taxi Trip Duration.

3. Shashank Badre. December 03, 2017. Exploratory data analysis on Green Taxi. Retrieved from:http://rstudio-pubs-static.s3.amazonaws.com/326454_0e4d6355b75a4578bebac6cd99cc319f.html

4. Chih-Ling Hsu. May 14, 2018. Analyze the NYC Taxi Data. Retrieved from: https://chih-ling-hsu.github.io/2018/05/14/NYC#q1-which-regions-have-most-pickups-and-drop-offs

5. Willy Sebastian. June 2, 2018. New York Taxi Trip Analysis. Retrieved from: https://rpubs.com/willyarrows/NYCTaxiTripsAnalysis

6. Xuefeng Peng, Yiming Pan, and Jiebo Luo, "Predicting High Taxi Demand Regions Using Social Media Check-ins," Special Session on Intelligent Data Mining, IEEE Big Data Conference, Boston, MA, December 2017.